DOCUMENT RESUME

ED 409 376 TM 026 956

AUTHOR Lunz, Mary E.; O'Neill, Thomas R.
TITLE A Longitudinal Study of Judge Leniency and Consistency.
PUB DATE Mar 97
NOTE 30p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Higher Education; *Interrater Reliability; Item Response Theory; *Judges; Longitudinal Studies; *Medical Education; *Pathology; *Scoring
IDENTIFIERS *FACETS Computer Program; Histology; Leniency (Tests); *Rasch Model

ABSTRACT

This retrospective longitudinal study was designed to show grading leniency patterns of judges within and across clinical examination administrations. Data from 17 different administrations of the histology examination of the American Society of Clinical Pathologists over 10 years were studied. Over the 10 years there were 4,683 candidates and 57 judges, of whom 41 provided data. Multifacet Rasch model techniques and the FACETS program were used to build a benchmark scale and then anchor subsequent administrations. Results show that judges vary in their levels of leniency, and that a judge is usually consistent in the application of his or her level of leniency across examination administrations. An appendix describes the FACETS model. (Contains 2 tables, 6 figures, and 10 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 409 376

A Longitudinal Study of Judge Leniency and Consistency

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY
Mary Lunz

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Mary E. Lunz

Thomas R. O'Neill

American Society of Clinical Pathologists

Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL, 1997

TM026956

BEST COPY AVAILABLE

A Longitudinal Study of Judge Leniency and Consistency

Abstract

This retrospective longitudinal ten study was designed to show grading leniency patterns of judges within and across clinical examination administrations. The multi-facet Rasch model techniques were used to build a benchmark scale and then anchor subsequent administrations. Results show that judges vary in their level of leniency and that most judges are consistent in the application of their level of leniency across examination administrations.

A Longitudinal Study of Judge Leniency and Consistency

Introduction

The purposes of performance examinations, of which oral and clinical examinations are samples, were discussed by Shepard et al (1996) and include the following: 1) enhance validity by representing a full range of desired outcomes; 2) preserve the complexity of knowledge and skill; 3) represent contexts in which the knowledge and skills are applied; and 4) adopt modes of assessment that enable candidates to show what they know. Linn et al (1991) in discussing validity indicated that performance assessments should transfer to specific tasks, maintain the cognitive complexity of the problems and represent content quality, and comprehensiveness. These comments emphasize the importance of structuring the content and format of performance examinations such that they represent the relevant skills and abilities needed to be deemed competent.

There is another aspect of performance assessments that must be considered, namely the individuals who assess or judge the quality of the candidate's performance. The ability of judges to make assessments seems to be taken for granted, while the communication literature (e.g. Tubbs and Moss, 1974) offers many substantial reasons why all judges are different, even if they undergo the same training and have comparable professional experience. Thus,it seems extremely important to understand and document whether, in fact, individuals who judge candidate performance follow the same or different grading patterns within and among candidates, within and across examination administrations. Understanding judge grading patterns is necessary to

have confidence in the assessment results (Camara and Brown, 1995) and to put holistic or composite scores into perspective (Reckase, 1995).

The judge is, of course, a critical part of clinical or oral examinations. However, little is known about the short term consistency and leniency of judges, and virtually nothing is known about the long term consistency and leniency of judges. Judge performance across administrations has rarely, if ever, been considered in empirical studies, partially because of the inter-judge reliability within administrations was considered sufficient evidence of reliability, and partially because data and methods for tracking judge performance across administrations were not available. Recent advances in psychometric methods have made tracking judge leniency possible. The Rasch measurement model (1960/1980) was extended by Linacre (1989) to the multi-facet Rasch Model (MFRM) which calibrates judge leniency and item difficulty estimates, then accounts for the impact of judge leniency, item and task difficulty, before a candidate ability estimate is calculated. The focus of this study is judge grading patterns across ten years of clinical examination administrations.

The Board of Registry of the American Society of Clinical Pathologists has administered a practical examination in histology for many years. During the last ten years, MFRM analysis methods have been used. Consequently, data were available for constructing a retrospective longitudinal study of judge performance during that period of time. Because data are from 17 different examination administrations, and different groups of judges graded in different years, there is a great deal of missing data. However, there is enough data to observe patterns of judge consistency among administrations even when there are missing grading sessions for individual judges. This is a retrospective study designed to obtain empirical information about patterns of

judge leniency over a long period of time. These judges were qualified experts in their field and trained to use the examination grading scale which was the same for all examination administrations.

There were two goals for this longitudinal study: 1) to identify the individual leniency patterns of judges across examination administrations, and 2) to identify consistency of judge leniency across administrations. Leniency is defined as the expectations imposed by the judge when evaluating candidate performance. Previous research (Lunz, Stahl, and Wright, 1996; Lunz, Stahl, and Wright, 1994) has demonstrated that judges impose different levels of leniency, but are usually fairly consistent in their application of that level of leniency among candidates within administrations, and across administrations.

Methods

Instruments

The histology clinical examination has four facets: 1) candidates, 2) judges, 3)slide-projects, and 4) tasks. Over the ten years there were 4,683 candidates, 57 judges, and 53 slide projects. The rating scale used to grade the three tasks (tasks = processing, microtomy and staining) for each slide-project remained the same during the period of this study. There was also sufficient overlap among judges, slide-projects and tasks to link the examination administrations into a benchmark scale. The multi-facet Rasch model used to analyze the data is presented in Appendix 1. The analysis of candidates, judges, projects and tasks are based upon the judges' ratings of the projects and tasks.

The ordering of the candidates, projects, judges, and tasks on a linear scale provides a frame of reference for understanding the relationship of the facets of the examination. It makes it possible to observe estimated candidate ability from highest to lowest, estimated item difficulty from most to least difficult, estimated judge leniency from most to least and estimated task difficulties from most to least difficult.

In the study, candidates were graded on three tasks for each slide project. Three judges scored portions of the candidate's performance, using a linked rotational pattern. In each administration, as well as across administrations, candidates took different examination forms because they were graded by different sub-groups of judges, often on different sub-groups of slide-projects. To make the candidate ability estimates comparable, the influence of the various facets must be systematically accounted for so that differences in candidate ability can be measured accurately and without contextual bias. This was accomplished using the MFRM.

In multi-facets analysis, candidate ability measures are estimated after the judge leniency and project difficulties encountered by the candidate are calibrated and equated to a benchmark scale. By placing all candidate performances on a scale, a comparable standard can be implemented for all candidates, even when the particular facet elements (e.g. judges) vary.

Benchmark Scale and Administrations

Data from 17 administrations were pooled and analyzed, thus placing them all on the same scale, called the benchmark scale. The FACETS program (Linacre, 1994) was used to calibrate candidates, judges, slide-projects, and tasks on the benchmark scale. To construct the benchmark scale the data from all 17 administrations were pooled and analyzed together. This

was possible because there was sufficient overlap of judges, slide projects and tasks across administrations to pull all facets onto the same scale. Administrations started in February, 1987 (labelled 2/87) and continued semi annually through May, 1996 (5/96). (Note: The first number indicates the month, the second number the year of the administration).

After the benchmark scale was established, individual examination administrations were analyzed separately. The difficulty estimates for the slide-projects and the tasks, as well as, the candidate ability measures from the benchmark scale were used to anchor the individual examination administration analyses (see Figure 1). Thus the non-anchored facet across administrations was judge leniency, so differences in judge leniency when the other facets were anchored, could be tracked across administrations.

A total of 57 judges participated in at least one of the 17 administrations; however, 16 judges graded in only one examination administration. The performance of these 16 judges is not reported, because it was not possible to observe their consistency among administrations. On average, judges graded in six administrations. Different subsets of judges graded during each administration. However, there were always some judges that overlapped among administrations. Figure 2 shows examples. Judges 18 and 40 overlapped in administration 8/92 while judges 40 and 57 overlapped in administration 5/94, so judges 18 and 57 are connected through their common link with judge 40.

Over the ten years, 53 different slide-projects were used, although a subset of only 15 slide-projects was used during any single examination administration. There was adequate overlap of slide projects to link administrations through subsets of common slides. The same three tasks were used to grade every slide-project across all administrations, thus providing

complete overlap for that facet and an important link among administrations. During the 10 year period, 4,683 candidates attempted the examination. The number of candidates attempting the examination in a single administration ranged from 168 to 385 with 275 candidates being the average. Candidates that overlapped administrations were considered to be different people because they had hopefully taken some action to improve their ability before attempting the examination a second time.

The leniency estimates of the judges for each administration were tracked when all other facets are anchored to the benchmark scale. The multi-facet judge leniency estimates were transcribed to scaled scores where 0 points was the most lenient judge and 100 was the most severe judge. The goal was to identify judges' patterns of leniency and consistency across administrations. Since all administrations were anchored to the benchmark scale, it is possible to observe differences in judge grading leniency across administrations.

Several graphs were developed to show key grading patterns. All graphs have the same vertical and horizontal coordinates, so it is possible to compare judge patterns across administrations.

Results

Mean candidate ability estimates across administrations were verified as not significantly different. This means that overall differences in candidate ability do <u>not</u> account for differences in judge leniency across administrations. Table 1 shows the mean leniency of the subgroup of judges for each examination administration. The range of judge leniency means, across administrations, was 35 to 52 scaled score points, which is a 17% difference overall. The range column shows

that the leniency levels of individual judges varied within administration (e.g. 2/89, range 17-100 scaled score points). Thus, the particular sub-set of three judges assigned to grade a candidate's performance could have a significant impact on the candidate's outcome. Table 2 shows the 43 judges in leniency order and the total number of examination administrations they graded. The majority of judges are in the moderate range; however, there is a substantial difference in the leniency represented by a scale score of 31 and 59 (30% difference). The graphs provide additional detail regarding leniency patterns of selected judges.

Most judges graded in some administrations and skipped others. Some judges graded many sessions, while others graded few. Some judges varied, from their expected grading pattern in one or more administrations, while others were extremely consistent. The following graphs are representative of patterns of judge grading across administrations.

Figure 3 shows the comparison of a severe and a lenient judge. The mean leniency of judge 46 was a scaled score of 64 points, while the mean leniency of judge 5 was a scaled score of 27 points. Each of these judges graded in 13 administrations and varied within 20 points of their average leniency across examination administrations. On the 5/96 administration, judge 5 was more lenient than in previous administration. Judge 46 was most severe in the 2/89 administration, the second administration in which she graded.

Figure 4 shows a consistent and an inconsistent judge. Each of these judges graded at 10 of the 17 administrations, so they skipped some administrations during the 10 year period. The average leniency of both of these judges was 43 scaled score points; however, judge 7 tended to vary at each administration, while judge 6 showed little variance after the first several examination administrations, even when administrations were missed.

Figure 5 shows that judges can be consistent, even when they do not grade in consecutive examination administrations. Judge 38 graded three consecutive administrations, then missed four consecutive examination administrations, but stayed within a 10 point leniency range. Judge 1 graded in one administration, then missed four administrations, then graded one administration, then missed four administrations, but remained within a 10 point leniency range. Notice that while these judges did not grade in every session, there is a certain amount of overlap because they graded in some of the same administrations.

Figure 6 shows two judges who moved from relatively severe to relatively lenient in the administrations they graded. Note that some sessions were missed, but the pattern of becoming more lenient is fairly obvious for these judges.

Discussion

These results confirm that judges are different from each other, but are generally consistent in their personal level of leniency across examination administrations, regardless of the candidate performances graded, or the slide-project graded. Even though the judges receive training and practice sessions before each grading session, they maintain their level of leniency across years, even when they do not grade consecutive examination administrations. In fact, one judge maintained a relatively uniform level of leniency across 15 administrations with 15 training sessions, spanning ten years.

Whether judges are, in fact, consistent or inconsistent across administrations, these results confirm the necessity of accounting for the differences in judge leniency within administrations. Table 2, shows that the candidate graded by judge 35 (leniency = 23 points) has a much higher

probability of earning a passing score than the candidates graded by judge 40 (leniency = 76 points) even if the scores these judges give are highly correlated. The inter-judge correlation shows that the judges follow the same pattern of awarding point values, it does not mean that they give similar value to the same candidate performance (see Lunz, 1992). It is important to remember that different subgroups of judges graded at each examination administration. Table 1 shows that mean judge leniency across administrations is relatively comparable (scaled score range of 35 - 52 points), even though the particular group of judges that graded in an administration varied. A single candidate does not interact with all of the judges, and the leniency of the particular subgroup of judges with whom the candidate interacts may vary substantially. In addition the range of judge leniency varies across examination administrations. Administration 2/89 showed the largest range of leniency scores (17 - 100). Most sessions had approximately 50 scaled score points of variance in judge leniency.

As Camara and Brown (1995) indicate, the purpose of the examination is to 1) make decisions about candidates, 2) aid instruction, and 3) provide evidence for accountability. Therefore, the decisions should be as reliable and reproducible as possible. Understanding judge grading patterns is an important issue in achieving this goal. These results provide useful information regarding judge grading patterns across administrations.

The limitation of this study is that it is retrospective, although it spans 10 years and 17 examination administrations. These examinations occurred from one to ten years ago, and these results were not available to assist with the annual pass/fail decision making process. While seeing the patterns may be interesting and helpful for understanding how judges grade examinations, the question is, how can this information be applied prospectively to insure that

candidates within and across examination administrations have a comparable opportunity to pass.

History yields several concepts that may be useful. First, there is sufficient consistency in the performance of most judges across administrations, and usually sufficient overlap between examination administrations to set up an equating design that would place more than one examination administration on the same scale. To construct the retrospective longitudinal study, all data were pooled and analyzed; however, examination administrations could be linked together through common judges, projects and even candidates after each administration and placed on a benchmark scale prospectively. Second, the retrospective study shows that clinical examination data from different examination administrations can be located on a benchmark scale. Similar techniques can be used to construct a benchmark scale prospectively. Once a benchmark is established, a criterion standard, that would apply to all candidates, within an administration, and then be carried forward to subsequent administrations could be established. Thus candidates would be required to meet the same criterion standard regardless of when they took the examination.

There are many issues such as judge training, the development and meaning of the rating scale, and the types of projects that are graded that contribute to constructing a common benchmark scale. In this longitudinal study the judges were trained using comparable techniques for each semi-annual administration, the tasks that were graded and the definition of the rating scale remained the same, and projects were in a defined domain. The groups of judges varied for each administration, the groups of projects varied for some administrations and certainly the candidates were different. However, the commonalities that were consciously maintained during the 10 years served to establish a sufficient amount of overlap to complete the longitudinal study.

If this can be accomplished retrospectively, with a little planning it can certainly be accomplished prospectively as well.

References

Camara, W and Brown, D. (1995). Educational and employment testing: Changing concepts in measurement and policy. Educational Measurement: Issues and Practice, 14, 1, 5-11.

Linacre, J. (1989). Many-Faceted Rasch Measurement. Chicago, IL: MESA Press.

Linacre, J. (1994). FACETS, a computer program. Chicago, IL: MESA Press.

Linn, L. Baker, E. and Dunbar, S.B. (1991). Complex, performance-based assessment: expectations and validation criteria. Educational Researcher, 20, 8, 15-21.

Lunz, M.,, Stahl, J. & Wright, B. (1996). The invariance of judge severity calibrations in Objective Measurement Theory into Practice, Vol. 3. (eds. Engelhard and Wilson) New Jersey: Ablex.

Lunz, M. (1992). New ways of thinking about reliability. Professions Education Researcher Quarterly, *13*(4), 16-18.

Lunz, M., Stahl, J and Wright. (1994). Interjudge reliability and decision reproducibility. Educational and Psychological Measurement, 54, 4, 913-925.

Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.

Reckase, M. (1995). Portfolio assessment: A theoretical estimate of score reliability. Educational Measurement: Issue and Practice, 14, 1, 12-14.

Tubbs, S. and Moss, S. (1974). Human Communication: An Interperson Prospective. New York: Random House.

Table 1

Average Judge Leniency Across Administrations

Administration	N Judges	Mean Leniency* of Judges	SD*	Range*
2/87	12	39	14	21-70
2/88	10	45	14	24-72
8/88	19	52	13	24-75
2/89	17	50	27	17-100
8/89	17	48	14	25-74
2/90	15	47	11	27-71
8/90	19	47	12	20-67
8/91	17	41	12	21-64
2/92	14	44	17	11-77
8/92	17	43	19	14-79
2/93	11	50	19	23-99
8/93	18	47	16	18-77
5/94	16	47	15	26-77
11/94	17	47	15	26-73
5/95	11	45	21	11-74
11/95	16	35	17	6-62
5/96	11	35	17	1-67

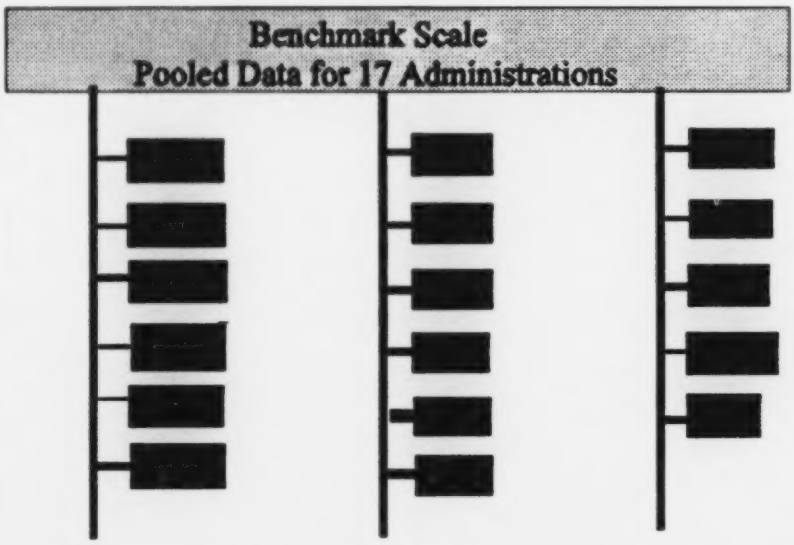
*Reported in scaled scores 0 - 100

Table 2
Judges in Leniency Order

Judge Number	Mean Scaled Score*	Number Administrations Graded	Category
35	23	9	
51	26	2	
5	27	13	Lenient
22	29	3	
16	31	4	
30	32	2	
62	33	13	
9	35	13	
13	36	3	
34	37	7	
18	37	5	
29	37	4	
26	38	5	
25	38	2	
1	39	4	
52	40	4	
38	40	4	
65	40	2	
4	43	15	
7	43	12	Moderate
6	43	11	
2	45	9	
27	46	6	
20	47	2	
32	47	2	
50	47	6	
19	51	14	
8	51	5	
12	52	5	
10	52	3	
36	53	9	
3	54	9	
59	56	3	
15	58	3	
53	59	2	
23	59	10	
57	64	5	
46	64	13	Severe
63	67	2	
40	76	5	
45	78	2	

* Reported in equated scaled scores 9 - 100
Low score = lenient judge
High score = severe judge

Figure 1
Benchmark Scale and Anchored Annual Administrations



All candidate ability estimates are on the same scale
All Slide-project difficulty estimates are on the same scale
All task difficulty estimates are on the same scale
All examiner lenience estimates are on the same scale

All examination administrations are equated because all slide-project, task, judges and candidates are calibrated to the same scale. Thus different examination administrations can be given and judge lenience can be tracked across administrations.

First number indicates the month, second number indicates the year of the administration.

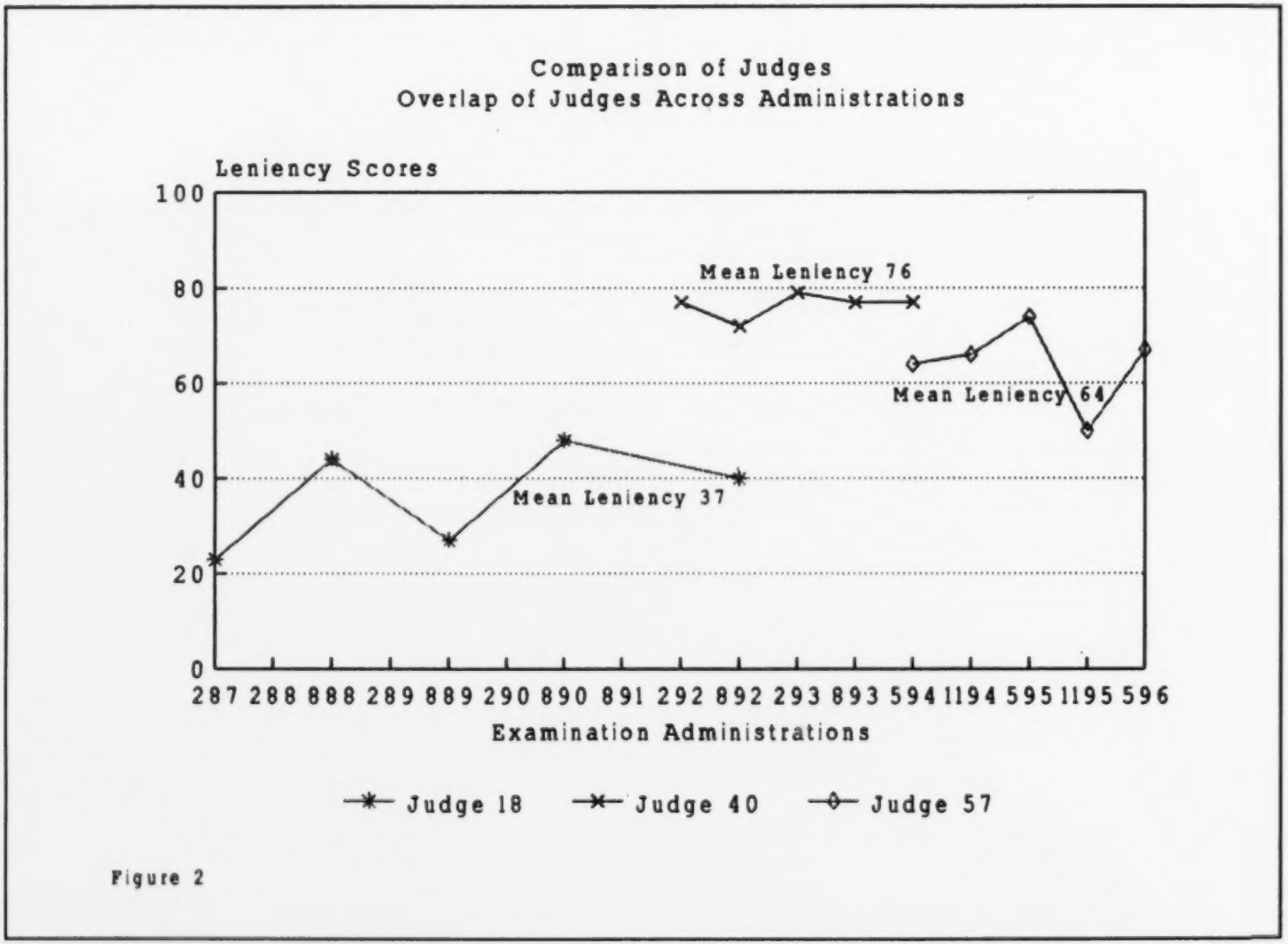

Comparison of Judges
Overlap of Judges Across Administrations
Leniency Scores
100
80
60
40
20
0
Mean Leniency 76
Mean Leniency 64
Mean Leniency 37
287 288 888 289 889 290 890 891 292 892 293 893 594 1194 595 1195 596
Examination Administrations
Judge 18
Judge 40
Judge 57

Figure 2

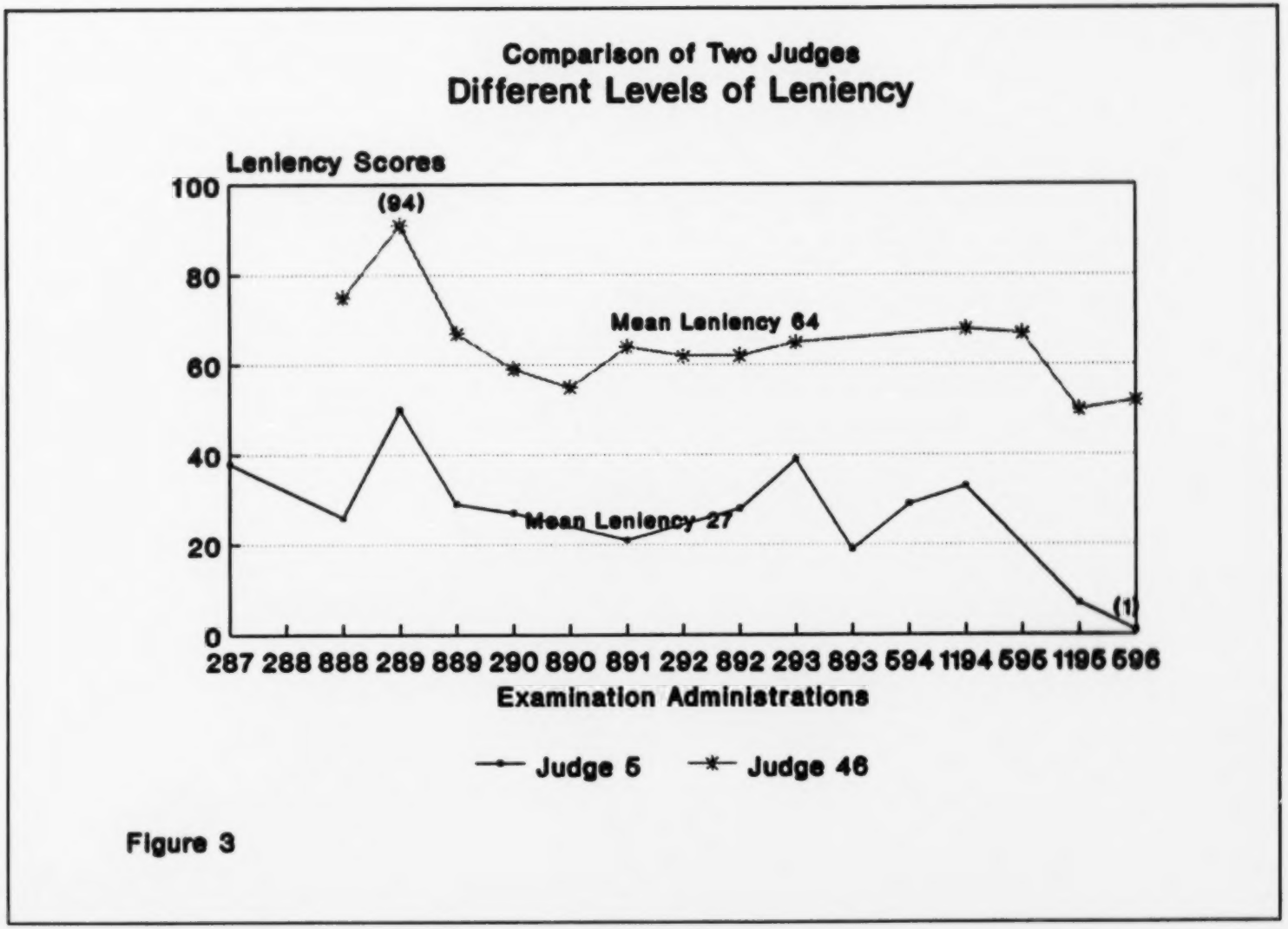

Comparison of Two Judges
Different Levels of Leniency
Leniency Scores
100
80
60
40
20
0
(94)
Mean Leniency 64
Mean Leniency 27
(1)
287 288 888 289 889 290 890 891 292 892 293 893 594 1194 595 1195 596
Examination Administrations
Judge 5
Judge 46

Figure 3

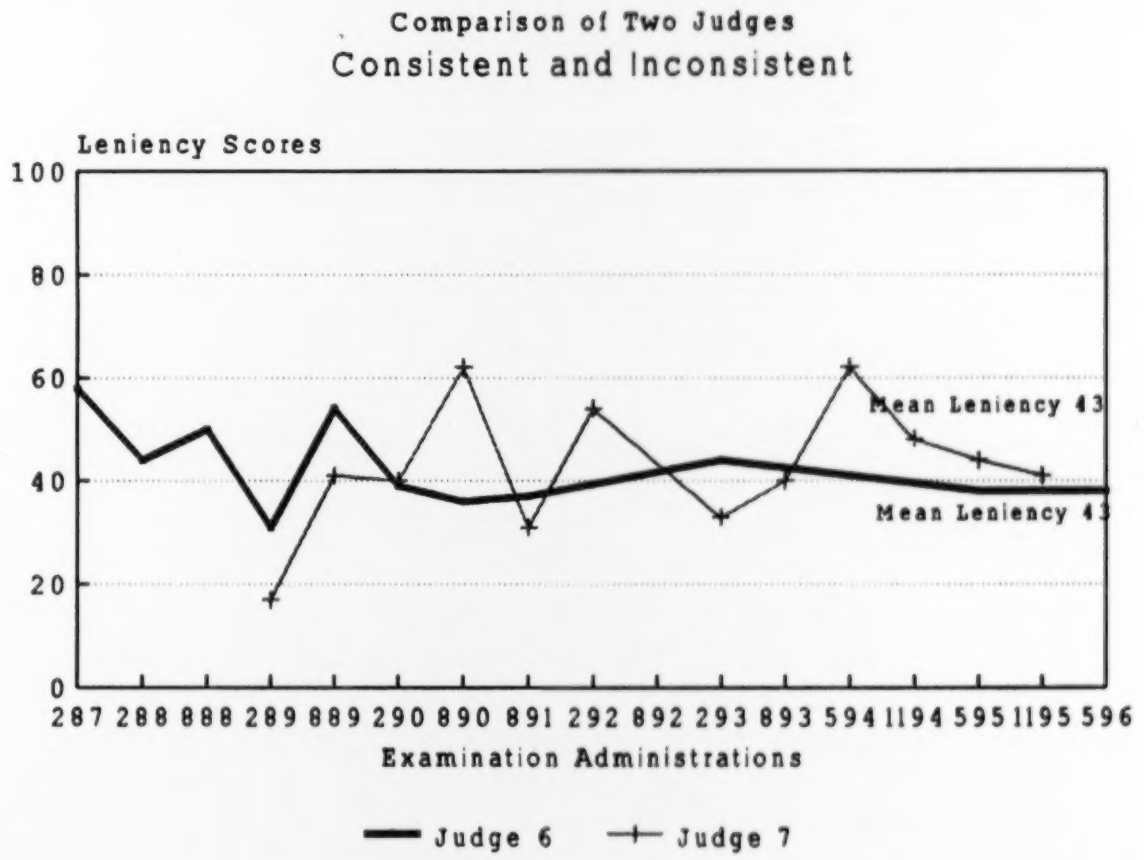

Comparison of Two Judges
Consistent and Inconsistent
Leniency Scores
100
80
60
40
20
0
Mean Leniency 43
Mean Leniency 43
287 288 888 289 889 290 890 891 292 892 293 893 594 1194 595 1195 596
Examination Administrations
Judge 6
Judge 7

Figure 4

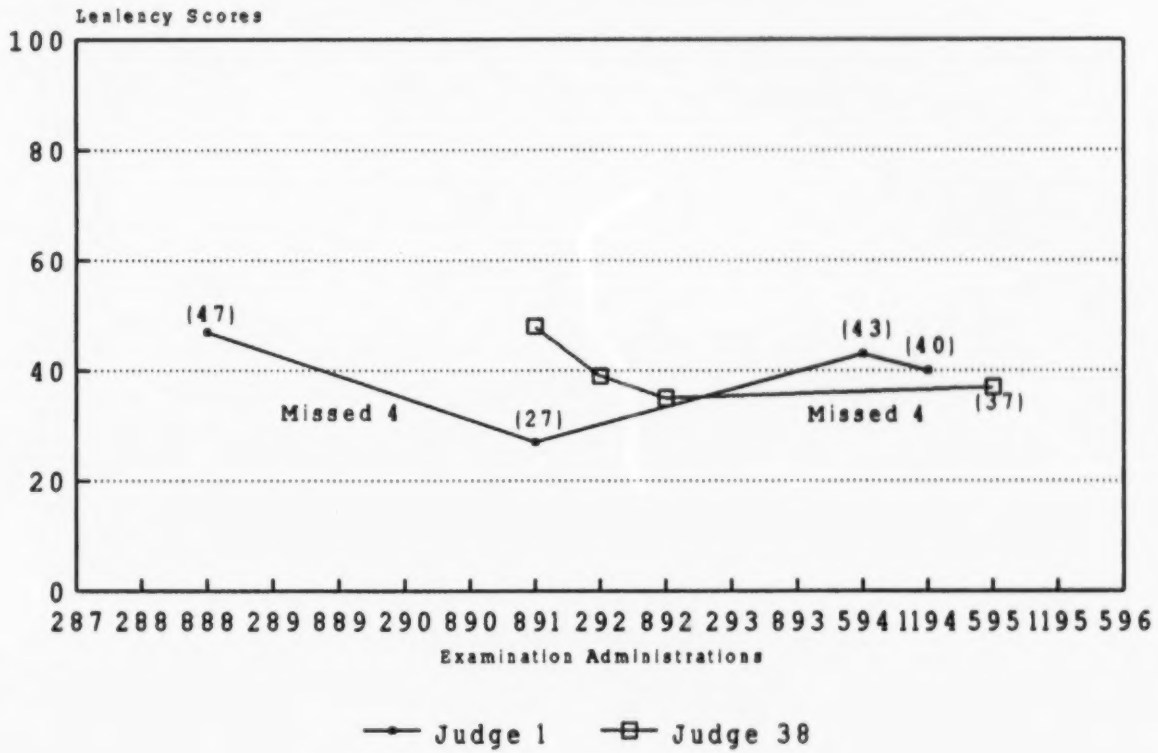

Comparison of Judges
Patterns Missed Administrations
Leniency Scores
100
80
60
40
20
0
(47)
Missed 4
(27)
(43)
(40)
Missed 4
(37)
287 288 888 289 889 290 890 891 292 892 293 893 594 1194 595 1195 596
Examination Administrations
Judge 1
Judge 38

Figure 5

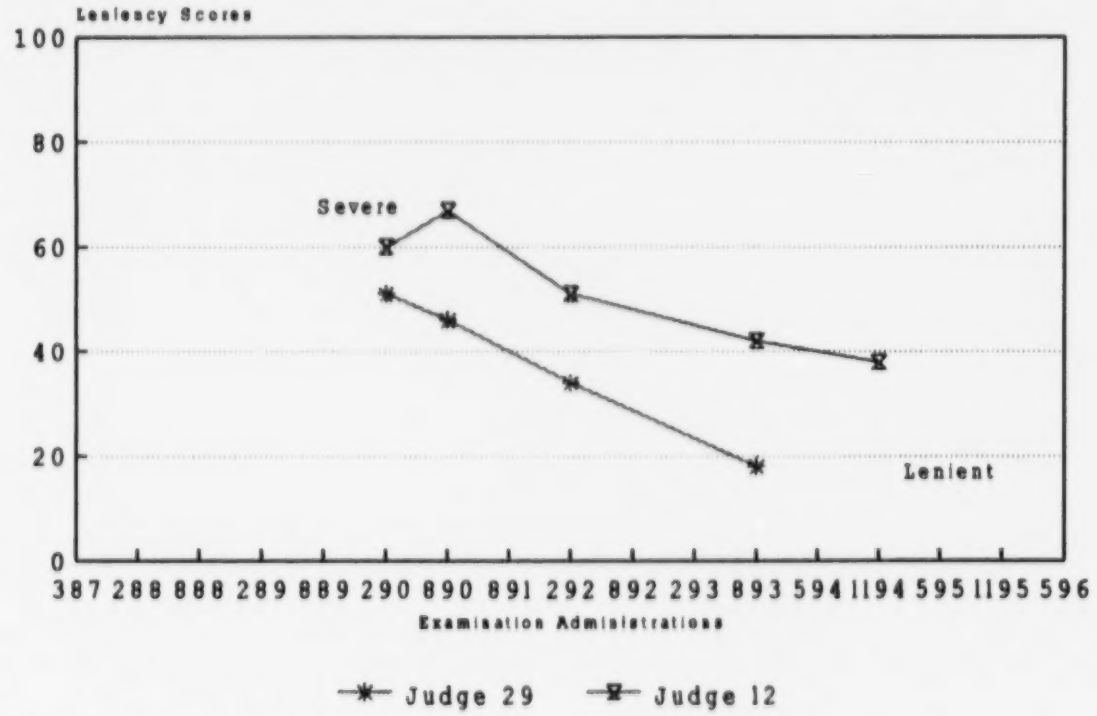

Comparison of Judges
More Severe to More Lenient
Leniency Scores
100
80
60
40
20
0
Severe
Lenient
387 288 888 289 889 290 890 891 292 892 293 893 594 1194 595 1195 596
Examination Administrations
Judge 29
Judge 12

Figure 6

Appendix 1

The Facets Model: background and explanation

The basic Rasch Model (Rasch, 1960/1980) is a mathematical representation of the person and item interaction. The log odds of a person answering a particular item correctly is modeled as:

$$\log(P_{ni}/(1 - P_{ni})) = (B_n - D_i)$$

where (P_{ni}) is the probability of answering the item correctly

($1 - P_{ni}$) is the probability of answering the item incorrectly

B_n is the ability of the person

D_i is the difficulty of the item

The probability of a correct response is a function of the difference between the ability of a person and the difficulty of the item. If a person's ability is greater than the difficulty of the item, then the probability of answering correctly is greater than 50%. If the difficulty of the item is greater than the ability of the person, the probability of answering the item correctly is less than 50%. The use of the logarithmic function in the equation transforms ordinal raw scores to a linear scale. The unit of measure is log-odds units or "logits" (Wright and Stone, 1979).

For analysis of assessments this basic Rasch model is extended to the multi-facet Rasch model (Linacre, 1989), so that facets for skill and item difficulty, evaluator leniency, and rating scale usage can be added to the equation. Leniency is the term used to encompass the factors that influence the way judge rate person performances. The difficulty of the item is the likelihood of responding correctly. Easy items represent tasks or general knowledge. More difficult items require more knowledge and skill or are more complex to perform. When a person is rated, the log odds of succeeding is modelled:

$$\log((P_{nxji})/(1 - P_{nmji})) = (B_n - T_m - C_j - D_i - R_k)$$

Where: (P_{nxji}) is the probability of performing the skill successfully

($1 - P_{nmjik-1}$) is the probability of performing the skill unsuccessfully

B_n is the ability of the person

T_x is the difficulty of the item

C_j is the leniency of the judge

D_i is the difficulty of the skill

R_k is the difficulty of rating step k compared to step k-1 on rating scale

The probability of a satisfactory performance is a function of the difference between the ability of the person and the difficulty of the skills after adjustment for the leniency of the judges and the difficulty of the projects. If the person's ability is higher than the difficulty of the project after adjustment for the difficulty of the skill and the leniency of the judge, then the probability of a satisfactory performance is greater than 50%. Conversely, if the difficulty of the project after adjustment for the difficulty of the skill and the leniency of the judge, is greater than the ability of the person, the probability of achieving a satisfactory performance is less than 50%.